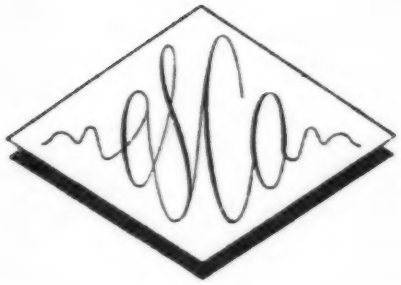
Proceedings of ESCA workshop

Speech Input/Output Assesment and Speech Databases



Noordwijkerhout, the Netherlands, 20-23 september 1989

DCIEM No. 89-P-43

TESTING THE EFFECTIVE VOCABULARY CAPACITY METHOD OF EVALUATING SPEECH RECOGNIZERS

J.M.E. van de Vegte and M.M. Taylor

Defence and Civil Institute of Environmental Medicine
P.O. Box 2000, 1133 Sheppard Ave West, North York, Ontario, Canada, M3M 3B9

ABSTRACT

The performance of isolated-word speech recognizers is typically measured using error rates. To obtain reliable error rate estimates for good recognizers can require many thousands of subject utterances, and, the better the recognizer, the more sensitive the error rate is to the actual challenge vocabulary and the skill of the talker. The Effective Vocabulary Capacity (EVC) is the maximum vocabulary that a recognizer can in principle handle at a given error rate. It relies on measures that are relatively independent of the challenge vocabulary and which require only tens or hundreds of test utterances. The EVC algorithm is tested for both synthetic and real recognizer data.

1. INTRODUCTION

Most methods of evaluating the performance of speech recognizers depend on the analysis of errors made when the recognizers are challenged by some standard test vocabulary. When the recognizers are good, they make few errors, and to compare two high-quality recognizers necessarily demands both a carefully chosen test vocabulary and a large number of challenge utterances. The Effective Vocabulary Capacity (EVC) method (Taylor, 1985; 1986; Vignault-Rasiulis et al., 1985) does not depend on the machine making errors, neither is it, to a first approximation, dependent on the challenge vocabulary. The EVC of a recognizer is the largest possible vocabulary that could be designed for the recognizer so as to give a specified (small) error rate.

The EVC method depends on the idea that a recognizer uses "templates" that can be located in some abstract space insofar as any two templates are separated by a "distance." An utterance produced by a talker can be assessed as being at some distance from all of the templates, and will be recognized as an instance corresponding to the nearest template, if it is not rejected by the recognizer. At the heart of the EVC method as currently implemented is a specially-developed multidimensional scaling algorithm (Vignault-Rasiulis et al., 1985) that operates under the assumption that the distances in question are Euclidean, an assumption that is probably not met by similarity or likelihood measures that can be extracted from realistic recognizers. This paper discusses tests of the robustness of the EVC method using simulated recognizers that produce (possibly) non-Euclidean distance measures.

The current EVC algorithm assumes that the analyst has access to several square distance matrices of which the rows represent utterances, and the columns represent templates. To form each such matrix, one utterance is spoken as an attempt to say each template, and its "distance" to each template is assessed by the recognizer. Neither the recognizer nor the analyst has any information about the distance from one template to another, about the distance from one utterance to another, or about the underlying coordinates associated with the templates and utterances. It is possible, from one such complete matrix, to recover the locations of templates and utterances in a multidimensional space whose axes have a well-defined orientation that maximizes the variance of the coordinate distributions on each axis in turn. The volume of the smallest multidimensional rectangle that contains all the templates and has its sides parallel to the axes of the space is known as the template volume. Across several distance matrices, the distribution of utterances around each template can be determined, and the average size of such distributions is known as the utterance volume, or error volume. The Effective Vocabulary Capacity of the recognizer is the ratio of the template volume to the error volume. EVC can, in principle, be computed from a single utterance per template, but more utterance sets are needed when the recognizer's similarity measures do not conform to the Euclidean distance assumption to discover and correct for the failures of assumption.

2. THE EVC ALGORITHM

The EVC algorithm addresses the problem of finding utterance (U) and template (T) coordinates such that (Distance (U,T))^2 matches a specified matrix of squared Euclidean distances, D. It relies on the techniques of metric unfolding (Schönemann, 1970; Gold, 1973) and matrix decomposition (Schiffman and Falkenberg, 1968). Metric unfolding allows location of two sets of points in a joint space, given Euclidean distances between elements from

distinct sets, and involves the calculation of a doubly-mass centred matrix C such that $-0.5\,C = U\,T$, where U is a (pxn) matrix representing p utterances, T is an (nxq) matrix representing q templates, and n is the dimensionality of the space. In addition, the row and column averages of the squared U-T distance matrix computed from the U and T coordinates must match those of the original D matrix. Schiffman and Falkenberg's work complements the metric unfolding of the D matrix by providing a principal components decomposition of $-0.5\,C$ as $G\,\beta\,H$, where G is the matrix of eigenvectors of $-0.5\,C$, β is the diagonal matrix of eigenvalues, and $H = \beta^{-1}\,G^{T}\,(-0.5\,C)$. (At this step, it is possible to recover, for Euclidean data at least, the underlying dimensionality of the space which produced the D matrix by noting the number of significant eigenvalues in the β matrix.)

Based on empirical trials, Schiffman and Falkenberg split the $G\,\beta\,H$ product into U and T matrices for p=q according to $U = G\sqrt{\beta}$ and $T = \sqrt{\beta}\,H$. The matrices G^{T} and H are orthonormal, so all of the variance in the utterance and template spaces is contained in the β matrix. This result implies an equal partitioning of the variance between these two spaces. However, the set of templates associated with a recognizer is assumed to be normally distributed in n-dimensional space and each template acts as the centroid of a cluster of normally distributed utterances. Therefore, the variance of the template coordinates in any dimension should be smaller than that of the utterance coordinates by an amount equal to the variance of each cluster of utterances around its own template. This re-partitioning of the variance is accomplished by iteration. The new U and T coordinates are given by $U = G\,\beta / \sqrt{\beta - \rho I}$ and $T = \sqrt{\beta - \rho I}\,H$, where ρ represents the positive constant obtained by iteration. Thus, re-partitioning the variance causes the template space to become smaller, and the utterance space, to grow.

Once the utterance and template coordinates have been recovered, the EVC of a recognizer may be computed in three steps: first, find the volume of the n-dimensional rectangle described by the outermost template coordinates along each principal components axis; second, use the central χ^2 distribution of the utterance-to-own-template (squared) distances in the original D matrix to estimate an error volume; and third, calculate how many error volumes can be accommodated by the machine's template volume.

3. TESTING THE EVC METHOD

Many tests of the EVC algorithm involve synthetic data, for which a theoretical EVC may be calculated. During the computation of principal components, a rotation of the original space occurs. This rotation affects the calculation of template spans as compared to the spans of the original templates, so in order to compare computed EVC to "theoretical EVC," the original template coordinates are also rotated by principal components before the theoretical EVC is computed.

3.1 Partitioning of Variance

Figure 1 demonstrates the effect of re-partitioning described above. Part A shows a set of original templates (x) in 2-space, along with 2 utterances (.) for each template. The boxes enclose the template (solid) and utterance (dashed) spaces. Part B depicts the EVC algorithm's reconstruction of the U's and T's when Schiffman and Falkenberg's equal partitioning is used. The x's are the templates recovered from the first set of utterances, the circles, from the second set. Note that the template and utterance spaces are comparable in size, and do not compare well with the original data. In part C, the EVC reconstruction resulting from iterative re-partitioning is shown. The template and utterance spaces now have the correct relative sizes (compare with part A).

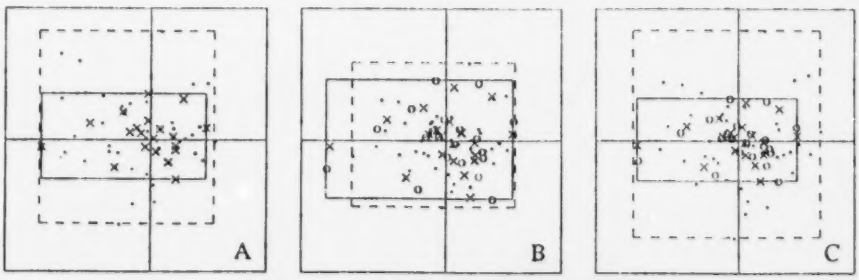


Figure 1 Re-partitioning the variance

3.1 Effect of Challenge Vocabulary

The EVC measure is designed to be relatively independent of challenge vocabulary. Figure 2 shows, for several vocabulary sizes, the mean EVC's obtained from ten random sets of templates drawn from a four-dimensional distribution. As the number of templates in the vocabulary increases, the EVC values indeed become more

consistent. For this synthetic data, figure 2 also shows the remarkable agreement of computed and theoretical EVC's with only the ten test utterances available for each template in this example.

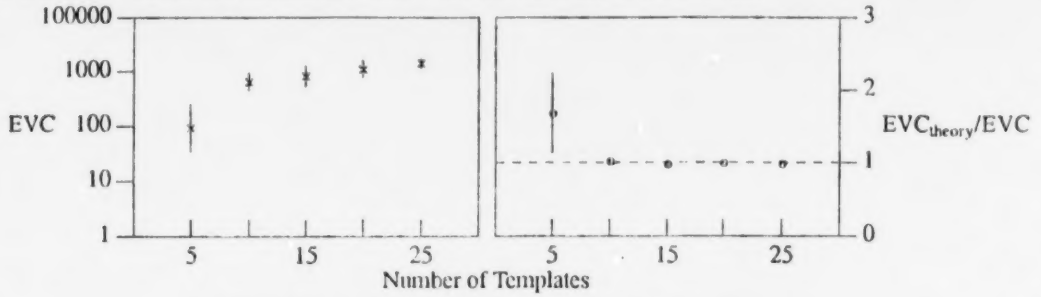


Figure 2 Effect of Vocabulary Size on EVC - Synthetic Data

The effect of vocabulary size on computed EVC for alphabetic-numeric data obtained from an actual recognizer is presented in figure 3. (Naturally it is impossible to compare computed EVC's with theoretical EVC values in this case.) For this recognizer, previous utterances act as templates, so the templates are not in fact the centroids of the utterance clusters, as is assumed by the EVC algorithm. Consistent EVC values are nevertheless obtained for randomly-selected vocabularies of twenty-five or more templates with only four utterances per template, computed dimensionalities ranging from 12 to 16. (Means are taken over ten template sets, except for the 30 and 36 template vocabularies, where only five sets are included.) In figure 3, two extra EVC estimates are shown for the fifteen template vocabulary. The circle represents the average EVC for a vocabulary hand-selected from the data to be as distinguishable as possible (e,h,l,o,q,r,w,x,y,z,1,4,5,7,0). The delta represents the average EVC for a vocabulary specially chosen to be as confusable as possible (a,b,c,d,e,g,j,k,p,q,t,u,v ?,3). It is clear that the random vocabulary selections give nearly the same EVC's as the vocabularies chosen for their distinguishability. The confusable vocabularies, on the other hand, give very low EVC estimates because they do not adequately represent the range of the working vocabulary.

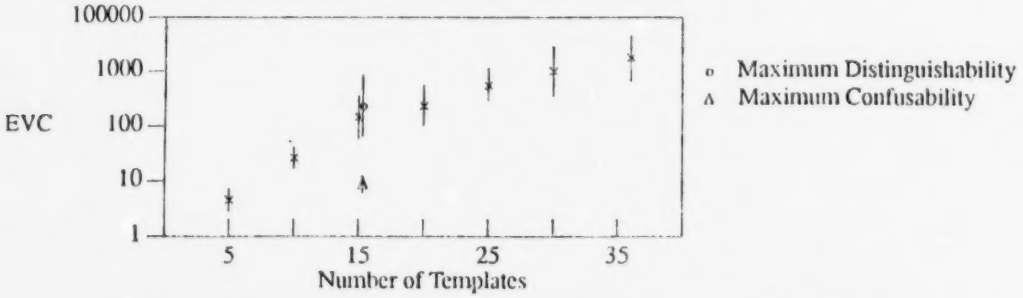


Figure 3 Effect of Vocabulary Size on EVC - Real Recognizer Data

3.2 Non-Euclidean Data

It is not reasonable to expect that raw "distance" data will necessarily be Euclidean. Therefore, the raw data is transformed, assuming that a power law relationship exists between the data in the D matrix and true Euclidean data. In a Euclidean space, variance is independent of (squared) distance, so an unbiased estimate of log(variance) plotted against log(distance squared) should produce a flat line. Such a plot is produced for distance matrices not known to be Euclidean and a regression line is fitted to the data. A power which is a function of the slope of the line is applied to the "distance" data, and the process is repeated. When the line reaches zero slope, the data is considered to have Euclidean characteristics. In particular, the estimator of variance is the difference between the non-central and central χ^2 means and variances: $\frac{var_{ij} - var_{ii}}{4(c_{ij} - c_{ii})} = \frac{2n\sigma^4 + 4\lambda\sigma^4 - 2n\sigma^4}{4(n\sigma^2 + \lambda\sigma^2 - n\sigma^2)} = \sigma^2$, where e_{ii}, e_{ij}, var_{ii} and var_{ij} are the means and variances, across all utterance sets, of the on- and off-diagonal entries of the D matrices, σ is the standard deviation of the utterance clusters, and λ is the χ^2 non-centrality parameter. The estimator of squared distance is provided by the off-diagonal means, the c_{ij}.

To test the transformation algorithm, matrices of Euclidean data were raised to known powers. Figure 4 presents the ratio of actual power to computed power for a collection of actual powers from 0.2 to 6.0, and for data of several different dimensionalities. The transformation algorithm consistently transforms the data to an effective

power between 0.8 and 1.2. For this range, computed EVC is compared to theoretical EVC (figure 5). Note that the average dimensionalities recovered by the principal components analysis (the integers annotating the solid lines) stray further from the correct values as the ratio of actual to computed power strays from one, producing generally inflated EVC estimates. When the analysis is artificially constrained to reflect the correct number of dimensions, underestimated powers give low EVC estimates and overestimated powers give high EVC estimates (the dotted lines in figure 5). In fact, to a first approximation, $EVC = (EVC_{theory})^{Ratio}$, for the dotted lines.

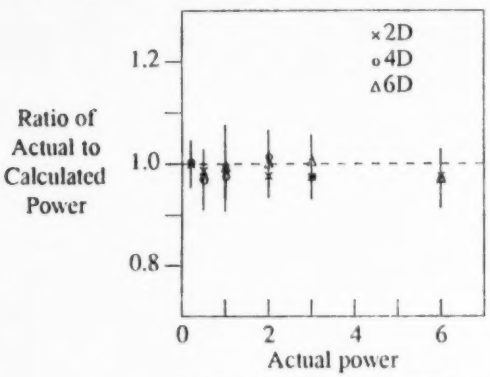


Figure 4 Data Transformation

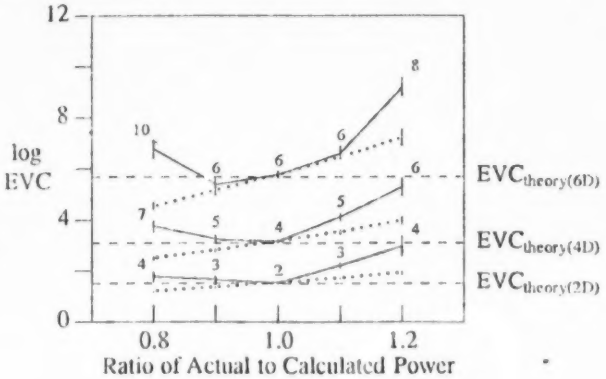


Figure 5 EVC Variability

4. CONCLUDING REMARKS

When the underlying templates are the centroids of utterance clusters and the data is Euclidean, the EVC algorithm produces, with only one to two hundred utterances, EVC estimates which are relatively independent of the size and composition of the challenge vocabulary, as long as the challenge vocabulary is reasonably representative of the working vocabulary. For some recognizers, it is not reasonable to assume that the templates are centroids of the utterance clusters. For example, when templates are actually previous utterances, EVC's are underestimated because the doubling of the number of degrees of freedom causes the size of the error volumes to be overestimated. When a cluster of templates is used in place of a single template, and the nearest template in each cluster is used as the target for the utterances, the EVC estimates tend to be high because the size of the error volumes is underestimated.

The errors inherent in using the transformation algorithm to transform non-Euclidean to Euclidean data are reflected in the EVC estimates for power data; the error in log(EVC) is approximately proportional to the ratio of actual to calculated power. The fact that the principal components analysis produces larger dimensionalities and hence larger EVC's as these errors increase complicates the problem. Data produced using city block metrics can also be transformed by the power law transformation, but the computed dimensionalities are often greatly overestimated, especially as the number of dimensions increases. Clearly, an independent estimate of dimensionality is needed for the transformation of these non-Euclidean measures.

ACKNOWLEDGEMENTS

Much of the programming and execution of the simulations was carried out by Doug Moseley as a student project.

REFERENCES

Gold, M.E. (1973) Metric unfolding: data requirement for unique solution & clarification of Schönemann's algorithm. Psychometrika, 38.

Schiffman, H. & P. Falkenberg (1968) The organization of stimuli and sensory neurons. Physiology and Behaviour, 3, 197-201.

Schönemann, P.H. (1970) On metric multidimensional unfolding. Psychometrika, 35.

Taylor, M.M. (1985) The Effective Vocabulary Capacity of Speech Recognizers: I. Problem. Canadian Acoustical Association.

Taylor, M.M. (1986) Issues in the Evaluation of Speech Recognition Systems. Journal of the American Voice I/O Society, 3, 39-67.

Vignault-Rasiulis, L., Jacobson, Z., Taylor, M.M. & J. van de Vegte (1985) The Effective Vocabulary Capacity of Speech Recognizers: II. Solution. Canadian Acoustical Association.